Molecular Phylogenetics and Evolution 111 (2017) 149–157

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Editor's Choice Article

The impact of GC bias on phylogenetic accuracy using targeted enrichment phylogenomic data

Silas Bossert [a,*], Elizabeth A. Murray [a], Bonnie B. Blaimer [b], Bryan N. Danforth [a]

[a] Department of Entomology, Cornell University, Ithaca, New York, USA
[b] Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, DC, USA

ARTICLE INFO

Article history:
Received 1 February 2017
Revised 6 March 2017
Accepted 24 March 2017

Keywords:
Gene tree incongruence
Ultraconserved elements
GC biased gene conversion
Gene trees
Sociality
Corbiculates

ABSTRACT

The field of sequence based phylogenetic analyses is currently being transformed by novel hybrid-based targeted enrichment methods, such as the use of ultraconserved elements (UCEs). Rather than analyzing relationships among organisms using a small number of genes, these methods now allow us to evaluate relationships with many hundreds to thousands of individual gene loci. However, the inclusion of thousands of loci does not necessarily overcome the long-standing challenge of incongruence among phylogenetic trees derived from different genes or gene regions. One factor that impacts the level of incongruence in phylogenomic data sets is the level of GC bias. GC rich gene regions are prone to higher recombination rates than AT rich regions, driven by a process referred to as "GC biased gene conversion". As a result, high GC content can be negatively associated with phylogenetic accuracy, but the extent to which this impacts incongruence among UCEs is currently unstudied. We investigated the impact of GC content on phylogeny reconstruction using in silico captured UCE data for the corbiculate bees (Hymenoptera: Apidae). The phylogeny of this group has been the subject of extensive study, and incongruence among gene trees is thought to be a source of phylogenetic error. We conducted coalescent- and concatenation-based analyses of 810 individual gene loci from all 13 currently available bee genomes, including 8 corbiculate taxa. Both coalescent- and concatenation-based methods converged on a single topology for the corbiculate tribes. In contrast to concatenation, the coalescent-based methods revealed significant topological conflict at nodes involving the orchid bees (Euglossini) and honeybees (Apini). Partitioning the loci by GC content reveals decreasing support for the inferred topology with increasing GC bias. Based on the results of this study, we report the first evidence that GC biased gene conversion may contribute to topological incongruence in studies based on ultraconserved elements.

1. Introduction

The field of phylogenetic analysis is changing rapidly due to the increasing accessibility of genome-scale data (Kjer et al., 2016; Lemmon and Lemmon, 2013; Wheeler and Giribet, 2016). Rather than analyzing phylogenetic relationships among organisms using a small number of genes obtained via PCR and Sanger sequencing, we can now evaluate phylogenetic relationships with many hundreds to thousands of individual loci. This change in scale transforms the tools that we use to resolve the tree of life, and opens new avenues to investigate deep evolutionary history. The inclusion of hundreds to thousands of individual gene loci in phylogenetic analysis, however, remains a challenge because of incongruence among phylogenetic trees of different genes or loci (Galtier and Daubin, 2008; Jeffroy et al., 2006; Nater et al., 2015), and the unequal distribution of topological conflict among clades (Wang et al., 2015). These issues of gene-tree/gene-tree and gene-tree/species-tree incongruence represent a persistent problem that does not necessarily disappear with genome-level datasets. In fact, phylogenomic datasets risk yielding wrong species trees with high confidence if significant incongruence is present, especially with concatenation based methods (Jeffroy et al., 2006; Kubatko and Degnan, 2007).

Topological discordance among gene and species trees can result from either methodological or biological factors (Rokas et al., 2003). Methodological factors are primarily conceptual or analytical limitations, such as the inadequacy of our current substitution models to accurately characterize sequence evolution (e.g., Kolaczkowski and Thornton, 2004; Roure and Philippe, 2011). In contrast, biological factors, such as incomplete coalescence in ancestral populations (incomplete lineage sorting, ILS) (Degnan

* Corresponding author.
E-mail address: sb2346@cornell.edu (S. Bossert).

http://dx.doi.org/10.1016/j.ympev.2017.03.022
1055-7903/

and Salter, 2005; Maddison, 1997; Pamilo and Nei, 1988), gene loss or duplication events (Galtier and Daubin, 2008; Goodman et al., 1979; Maddison, 1997), and introgression (Fontaine et al., 2015; Kronforst, 2008) can yield gene trees that are incongruent with each other and with the actual species tree, even though the genealogy is correctly inferred.

One important factor that impacts the level of incongruence in phylogenomic data sets is the composition of GC content. GC rich gene regions are linked to higher recombination rates than AT rich alleles, substantially driven by GC biased gene conversion (gBGC, e.g., Figuet et al., 2015; Kent et al., 2012; Lartillot, 2013). Experimental evidence revealed gene conversion to be biased towards the fixation of GC nucleotides instead of AT nucleotides when alleles are heterozygous (Lesecque et al., 2013; Mancera et al., 2008). This effect leads to unequal base compositional distributions in genomes due to higher fixation rates of GC in frequently recombining regions, so-called "recombination hotspots" (Duret and Galtier, 2009; Mancera et al., 2008). In turn, elevated recombination rates enhance incongruence among gene trees through increased probabilities of ILS, an effect which can be theoretically substantiated by speciation time and the effective population size (N_e) of given loci (Maddison, 1997; Pamilo and Nei, 1988). In fact, recombination rate and overall congruence were shown to be negatively correlated in different systems (Pease and Hahn, 2013) and ILS was specifically shown to increase with recombination frequency (Hobolth et al., 2011).

Generally, two different measures are used to identify frequently recombining, GC biased loci: Overall GC content and GC heterogeneity. By "GC content", we mean the overall percentage of GC across a range of aligned nucleotide sites (i.e., gene or locus). "GC heterogeneity" is the degree to which individual taxa deviate in GC content from the overall background level of GC content in the alignment (i.e., the variance of GC content among taxa). The latter measure is less widely used but has been shown to impact topological incongruence on phylogenetic reconstruction of the corbiculate bees using transcriptome data (Romiguier et al., 2015).

Phylogenomic data sets provide a strong framework for evaluating the impact of GC bias on gene-tree incongruence. One increasingly popular approach to generating genomic-level data sets of thousands of loci is hybrid-based targeted enrichment (Faircloth et al., 2012; Lemmon et al., 2012). Targeted enrichment methods allow one to effectively shortcut the difficult and labor-intensive challenges of whole-genome assembly and orthology identification for non-model taxa. Faircloth et al.'s (2012) approach uses in-solution sequence capture (sensu Gnirke et al., 2009) of large quantities of so-called "ultraconserved elements" (UCEs). These universal markers were shown to be superior to traditional multi-locus sequencing (Blaimer et al., 2015) and provide great applicability for degraded material, such as pinned museum specimens (Blaimer et al., 2016). Interestingly, UCEs are generally more AT-rich than GC-rich (Faircloth et al., 2012, Supplementary Fig. 2), suggesting that they may be less prone to the kind of gene conversion that can increase gene-tree incongruence in phylogenomic data sets. Both UCEs and selectively filtered AT-rich protein-coding phylogenomic data sets performed comparably well in resolving inconclusive nodes in mammal phylogenies (McCormack et al., 2012 and Romiguier et al., 2013, respectively). However, the impact of GC composition on incongruence among UCE loci is currently unstudied.

In this paper, we investigate the impact of GC content and GC heterogeneity on phylogeny reconstruction using in silico captured UCE data for the corbiculate bees (Hymenoptera: Apidae). Corbiculate bees include a total of 1019 species (Ascher and Pickering, 2016) placed in four tribes: Euglossini (orchid bees), Bombini (bumblebees), Meliponini (stingless bees), and Apini (honeybees). These four tribes are united by the possession of a highly modified hind tibia, which forms a concave, shiny structure (the corbicula) for carrying pollen and plant resins (Martins et al., 2014). Euglossini are typically viewed as solitary or communal but a number of studies have indicated that some species of *Euglossa* exhibit features typical of a more social bee, including multifemale nests, overlap of generations, and temporary division of labor in which some females forage and others guard the nest (reviewed in Cardinal and Danforth, 2011). The vast majority of the 260 species of bumblebees are primitively eusocial. Both Apini and Meliponini are advanced eusocial with obligate swarm founding and morphologically distinct castes.

The phylogeny of the corbiculate bees has been the subject of considerable study and establishing the relationships among the corbiculate tribes has been highly controversial. Previous morphological approaches supported a tree in which the advanced eusocial Meliponini and Apini formed a monophyletic group sister to Bombini, with Euglossini at the base of the tree (Fig. 1C; Cardinal and Packer, 2007; Plant and Paulus, 2016), suggesting a single origin of eusociality in corbiculate bees. However, numerous molecular studies (e.g., Cardinal and Danforth, 2011; Kawakita et al., 2008; Martins et al., 2014) have supported a very different view: that Bombini and Meliponini form a well-supported monophyletic group which is sister to Apini + Euglossini (Fig. 1A), thereby implying two separate origins of advanced eusociality. A recent phylogenomic investigation accounted for GC bias in coding sequences obtained yet another topology. Romiguier et al. (2015) obtained a tree in which Apini is sister to the clade of Bombini + Meliponini (Fig. 1B), with strong support when utilizing the non-homogeneous GG98 substitution model (Galtier and Gouy, 1998).

Besides the high degree of topological conflict, two aspects render the corbiculate bees an ideal group for investigating the impact of GC-bias on incongruence among ultraconserved elements. First, this group has been the focus of research on the genetic basis of sociality in bees and therefore there is an unusually rich data set of published and annotated whole genomes, with 10 newly published bee genomes in the last three years (Kapheim et al., 2015; Kocher et al., 2013; Park et al., 2015; Rehan et al., 2016; Sadd et al., 2015). These genomes can be effectively used for extracting in silico UCE data using the phyloinformatic pipeline Phyluce (Faircloth, 2016). Second, the main phylogenetic controversy in

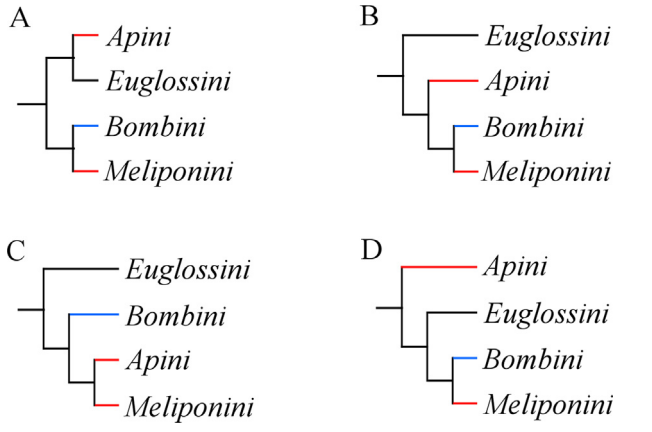

Fig. 1. Competing topologies of the corbiculate lineages. Topology A is favored by the majority of the single and multi-locus studies (e.g., Cardinal and Danforth, 2013; Kawakita et al., 2008), whereas topology B was revealed by a recent genomic study which accounted for GC bias (Romiguier et al., 2015). Topology C was inferred multiple times based on morphology (e.g., Cardinal and Packer, 2007; Plant and Paulus, 2016) but not by molecular data. Topology D is unlikely from the perspective of the evolution of eusociality but has been obtained using large data sets (Romiguier et al., 2015). Colored lines indicate eusociality of different degrees; red - advanced, blue - primitive. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

this group revolves primarily around a three-taxon statement involving Apini, Euglossini, and Bombini + Meliponini. Hence, we can apply quartet tree based evaluation methods, as the conflict is restricted to a single rooted triplet with three possible topologies (Fig. 1A, B, D). This study uses this publicly available data to gain first insights into base compositional effects on the performance of UCE data for phylogenomic inferences. We further utilize this new class of phylogenomic markers to explore the evolutionary relationships among the corbiculate bees.

2. Material and methods

2.1. In-silico sequence capture and processing

We used the Hymenoptera probe set designed by Faircloth et al. (2015) to extract UCEs from 13 publicly available bee genomes and UCE contigs of three additional bee taxa that were captured using this probe set (Table 1). Captured loci were extracted as fasta files with total up- and downstream flanking regions of 850 bp length. We called Mafft (Katoh et al., 2009) as implemented in Phyluce (Faircloth, 2016) to align the sequence data on the individual loci separately. We applied a 75% cutoff for data set completeness, meaning that each locus is represented by at least 75% of all taxa (i.e., a minimum of 12 taxa in total per locus). Capture statistics were calculated using Phyluce and Amas (Borowiec, 2016). Nucleotide variabilities of captured sequences in respect to their position to the UCE core were inferred with the Phyluce smilogram script and plotted with R 3.2.5 (R Core Team, 2016), utilizing the ggplot2 package (Wickham, 2009). An extensive description of the in silico capture methods and the subsequent processing are provided in the supplementary online information, including a step by step listing of the commands used (Supplementary data 1).

2.2. Locus selection and alignment filtering

Two alignments for concatenation-based phylogenetic methods were created. First, the loci of the 75% completeness matrix were concatenated without further processing ('Raw75p'). For the second dataset, we filtered the Raw75p dataset for the best performing loci based on the average bootstrap support of their respective gene trees. This filtering approach is primarily a workaround to bypass the extraordinary computational demands of analyzing extremely large, phylogenetic datasets. The same approach was used in recent UCE based studies (Blaimer et al., 2015; Branstetter et al., 2016). Here we extract the 100 highest scoring loci and compare their performance and base composition against the unprocessed dataset. Therefore we called RAxML 8.2.8 to calculate the best-scoring ML gene trees on each unpartitioned individual locus of the Raw75p dataset. A rapid bootstrapping analyses with 200 bootstrap replicates (-f a) was conducted, while specifying GTR+Γ as the substitution model for the bootstrap search. Utilizing the gene_stats R script used in Borowiec et al. (2015), we then inferred the individual average bootstrap support values for each gene tree. We subsequently screened for loci that yielded the 100 highest scoring ML gene trees based on their average bootstrap support and discarded all other loci ('Best100').

To account for base compositional effects among genes with different GC composition, two separate subsets of loci with different GC characteristics were created. We utilized the 810 loci of the Raw75p dataset and excluded the 10 worst performing loci based on their average bootstrap support to reduce the data set to a manageable 800 loci. First, we calculated the GC content for each locus individually. All loci were ranked by their overall GC content and divided into 20 groups of 40 loci, thereby letting each group represent 5% of all loci ('subset GC content'). This means that locus group 1 consists of loci with very low GC content, whereas locus group 20 is rich in GC. For the second subset, we calculated the variance of the GC content among taxa for each locus, as a measure for heterogeneity using a custom python script. The loci were ranked by increasing variance and again divided into 20 groups of 40 loci ('subset GC heterogeneity').

2.3. Phylogenetic inferences

To identify conflicting topologies among the individually processed datasets, species trees were inferred using both (1) concatenation and (2) gene tree methods.

2.3.1. Concatenation methods

For the concatenated multi-locus alignments, each locus was designated as a unique partition. We first executed best-scoring maximum likelihood (ML) tree searches and rapid bootstrap analyses with 1000 replicates for both concatenated datasets (Raw75p,

Table 1
Genomes and additional sequence data used in this study.

Genomes	Reference
Apis cerana	Park et al. (2015)
Apis dorsata	Unpublished; NCBI *Apis dorsata* Annotation Release 100, assembly accession GCF_000469605.1, part of the *Apis dorsata* Genome Sequencing Project (Accession PRJNA174631)
Apis florea	Unpublished; available on the webpage of the Baylor College of Medicine Human Genome Sequencing Center (https://www.hgsc.bcm.edu/arthropods/dwarf-honey-bee-genome-project)
Apis mellifera	The Honey Bee Genome Sequencing Consortium (2006, 2014)
Bombus impatiens	Sadd et al. (2015)
Bombus terrestris	Sadd et al. (2015)
Dufourea novaeangliae	Kapheim et al. (2015)
Eufriesea mexicana	Kapheim et al. (2015)
Habropoda laboriosa	Kapheim et al. (2015)
Lasioglossum albipes	Kocher et al. (2013)
Megachile rotundata	Kapheim et al. (2015)
Melipona quadrifasciata	Kapheim et al. (2015)
Ceratina calcarata	Rehan et al. (2016)
TRINITY-assemblies	
Andrena sp.	Faircloth et al. (2015)
Andrena asteris	Faircloth et al. (2015)
Bombus pensylvanicus	Faircloth et al. (2015)

Best100) using RAxML. We applied a GTR+Γ model for both the tree searches and the bootstrapping phases.

Bayesian tree inferences (BI) for each dataset were carried out by running four independent replicates of each four Metropolis-coupled MCMC using ExaBayes 1.4.2 (Aberer et al., 2014). The chains ran for 1.000.000 generations and were sampled every 25 generations, yielding 160.000 trees per alignment. Branch lengths were linked among partitions and a relative burn-in of 0.2 was used. Convergence was evaluated by ensuring effective sample size values (ESS) of >200 for each parameter, potential scale reduction factors (PSRF) ranging close to one and average standard deviations of split frequencies (ASDSF) of ≤0.5%. Posterior distributions of tree topologies of the individual runs were condensed using the implemented *consense* command and parameter files were saved with *postProcParam*.

2.3.2. Gene tree methods

Species trees using the multi-species coalescent model were inferred with Astral-2 4.10.6 (Mirarab et al., 2014; Mirarab and Warnow, 2015), as the software can properly handle unrooted gene trees with missing taxa and polytomies. Since summary methods such as Astral estimate species trees based on the topologies of the underlying input gene trees, the species trees are prone to be biased by poorly resolved gene trees of uninformative loci. We therefore in turn revised all individual gene trees and collapsed nodes with bootstrap support values of ≤50 using TreeCollapserCL 4 (Hodcoft, 2016). We subsequently summarized the 810 trees of the Raw75p dataset using the heuristic algorithm and inferred local posterior probabilities (PP; Sayyari and Mirarab, 2016) to measure topological support, as this measure was shown to be of superior accuracy than multi-locus bootstrapping *sensu* Seo (2008). Additionally, the quartet support for the individual quartet trees was calculated using the score option.

2.4. Tracing base compositional bias

We adopted and modified the topological incongruence measure of Romiguier et al. (2015) to explore trends of gene tree incongruences based on base composition. We therefore summarized the ML trees of the individual loci of both subsets ('GC content', 'GC heterogeneity') for each group separately using Astral-2, and scored the resulting species trees against each individual gene tree of their respective group. Instead of the quartet distance measure (cf., Romiguier et al., 2015), we extracted the quartet support (as defined in Sayyari and Mirarab, 2016) for the tested topologies (q1 in Astral-2) of every scored quartet and plotted their distribution as a function of the GC content and heterogeneity groups, respectively. In contrast to normalized quartet support of entire species trees, this measure is more sensitive to incongruence among single quartet trees, as it reflects the relative support of all branches and not the total of all satisfied quartets. It should therefore be superior in detecting effects of gene tree heterogeneity in actual biological data sets, as incongruence is unlikely to be equally distributed among entire phylogenies (cf., Wang et al., 2015).

As the sister group relationship of bumblebees (Bombini) and stingless bees (Meliponini) is well established, the investigated topological conflict is restricted to a single quartet tree with three possible hypotheses (Fig. 1A, B, D). We therefore scored the collapsed 40 ML trees of each locus group of both subsets ('GC content', 'GC heterogeneity') against the inferred topology of the concatenation based methods (Fig. 2; corresponds to topology B in Fig. 1). Subsequently, the local PP of the corresponding quartet tree for each of the three possible hypotheses was extracted for every locus group individually. We plotted the support for each hypothesis as a function of increasing GC content or GC heterogeneity, respectively, and measured the association with simple linear regressions. We tested for discordance of $H_0 = 0$ with simple linear regression models.

3. Results

3.1. Sequence capture and locus filtering

Matching the probes to the extracted genome slices and contig assemblies, and subsequent trimming yielded in a total of 1132 shared UCE loci with a mean alignment length of 518 bp. The level

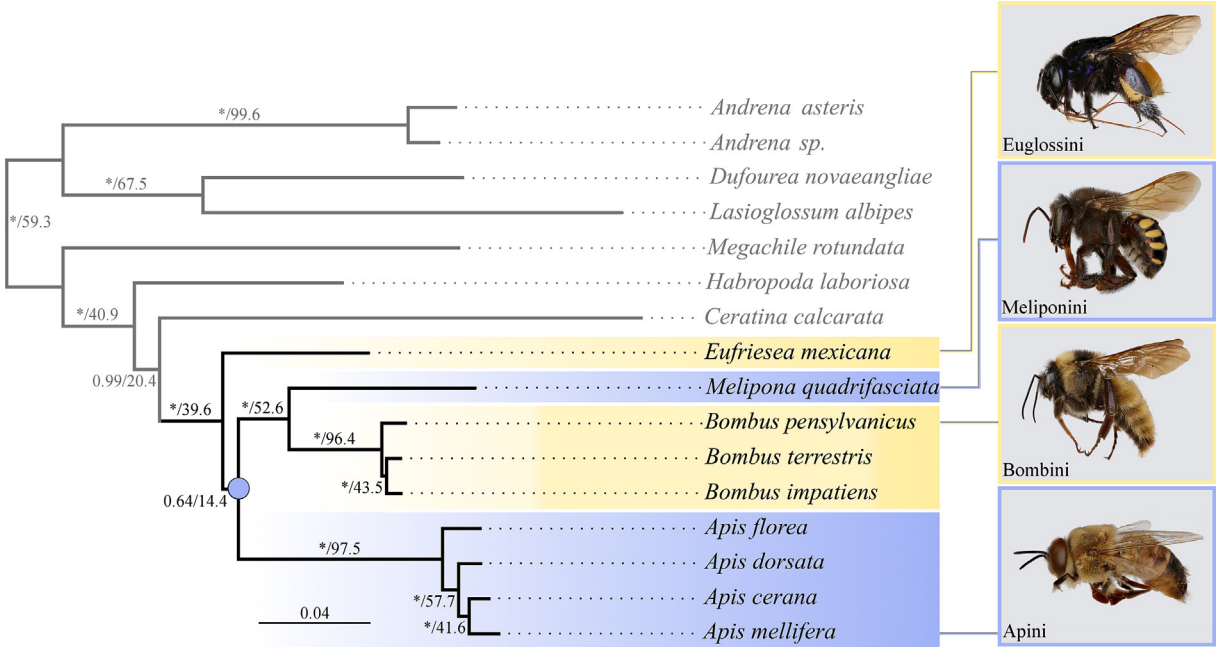

Fig. 2. Bayesian phylogram of the 'Best100' dataset. The scale bar indicates nucleotide substitutions per site. All nodes revealed Bayesian posterior probabilities and ML bootstrap support of 1 and 100, respectively. The branch values indicate the local posterior probability inferred by ASTRAL-2 and the actual quartet score. Asterisks correspond to 1.

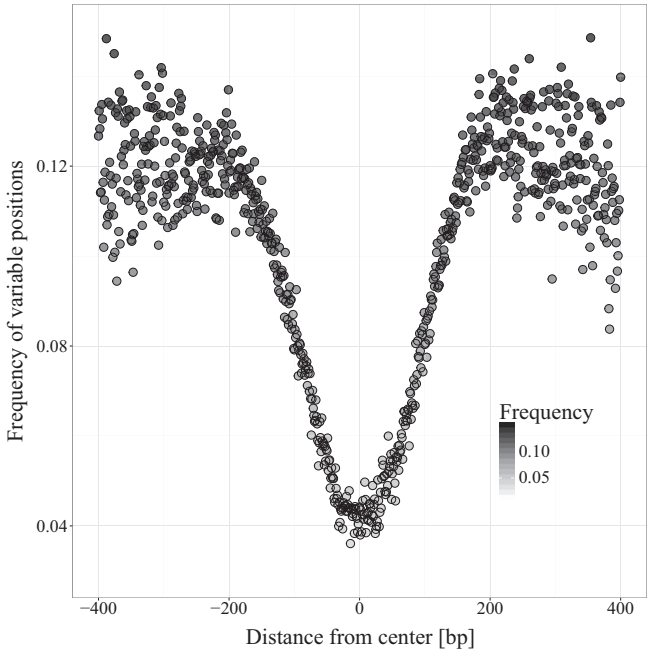

Fig. 3. Nucleotide variability of the captured ultraconserved elements. The center represents the core regions with low frequencies of variable sites. The frequency increases with the distance to the core region, displaying the greater variability of the flanks. Outliers are removed.

Table 2
Alignment statistics.

	Raw75p alignment	Best 100 alignment
Number of loci	810	100
Alignment length (bp)	419,250	73,434
Average locus length (bp)	518	734.34
Undetermined characters (%)	12.02	12.68
Parsimony informative sites	110,513	20,198
Overall GC content (%)	46.2	42.24
Mean GC variance among taxa and loci	5.48	9.04

of sequence variation of the UCE nucleotide positions in respect to their distance to the core revealed the expected 'smilogram'-pattern with highly conserved anchor regions and increasingly variable flanking regions (Fig. 3). An average of 894 loci was captured per taxon, with all taxa represented by at least 840 loci except for *Apis florea* with 613. Filtering the captured loci for a 75% completeness matrix reduced the number of loci for the subsequent alignment to 810, with 12.02% representing gaps or missing information. The amount of missing data of the Best100 alignment is comparable, whereas the average Best100 matrix is relatively longer, with a higher percentage of parsimony informative sites and is 4% lower in the overall GC content than the average Raw75p locus. The higher variation of the average GC content among taxa further reveals that the loci of the Best100 dataset are more heterogeneous in their GC composition (Table 2).

Overall, the GC content is roughly normally distributed among the captured loci, with the majority showing 45–55% GC after trimming. The homogeneity among taxa is less equally distributed with a large positive skew (Supplementary data 2 and 3). GC content and GC heterogeneity are strongly correlated ($R^2 = 0.167$, $p < 0.0001$), however, the few very heterogeneous loci are not necessarily rich in GC and vice versa (Supplementary data 4). Interestingly, the GC content of the most heterogeneous loci is unequally distributed among the included taxa. There is a trend that the corbiculate lineages are generally lower in GC content than other taxa. This is particularly apparent among the most heterogeneous loci (Fig. 4). Even if this trend is not shared among all loci, several of the most heterogeneous loci showed a clustering of the corbiculate lineages with lower GC content than all other taxa. Furthermore, several of these most heterogeneous loci produced well-supported species trees: The Best100 dataset includes 20 of the 10% most heterogeneous loci of the entire data set, whereas only 1 of the 10% GC richest loci falls into this category.

3.2. Coalescent vs. concatenated species trees and the inferred phylogeny

The concatenated species tree reconstructions of both the ML and BI approaches produced congruent topologies and maximum supported trees for both the 810 loci alignment ('Raw75p') and the Best100 locus set (Fig. 2). In contrast, the quartet supported gene tree recovered the same topology but reveals considerably lower confidence at the base of the corbiculate phylogeny, indicated by the fairly low local PP of 64. Only 14.4% of all induced quartet trees at this branch are actually congruent with the depicted topology, followed by 13.8% favoring the best alternative, which corresponds to topology D in Fig. 1. Notably, these values dropped from 36.3% and 36.2% (!), respectively, after collapsing nodes with bootstrap support values of ≤50. In contrast, all other quartet trees strongly support the concatenation- based trees. Besides the corbiculate taxa, the inferred phylogeny is in line with our current understanding of the higher-level relationships among the bees (Danforth et al., 2012). Moreover, it strongly supports the generic consensus tree of *Apis* sensu Koeniger et al. (2011), which was previously inferred using morphology and smaller scale molecular data analyses. It is further noteworthy that the phylogenetic position of the Xylocopinae (*Ceratina calcarata*) renders the remaining Apinae as paraphyletic (as expected; Danforth et al., 2012).

3.3. Effects of different locus selection

The GC sorted subsets of locus groups ('GC content', 'GC heterogeneity') had opposing effects on the general congruence among

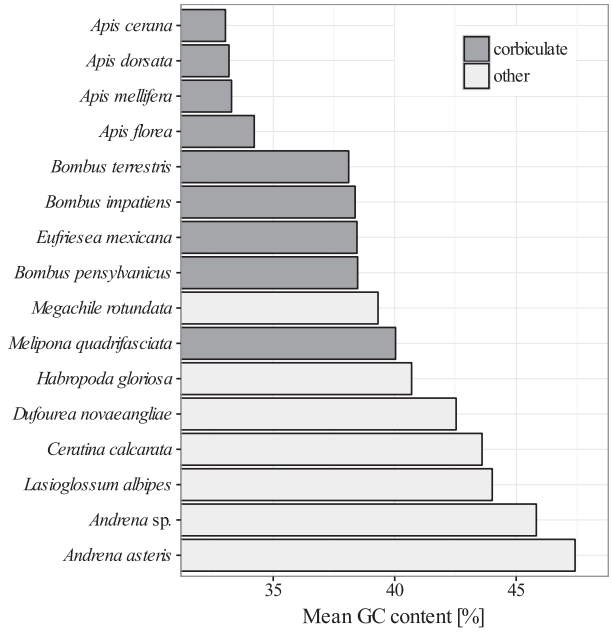

Fig. 4. Mean GC content of the 20 most heterogeneous loci.

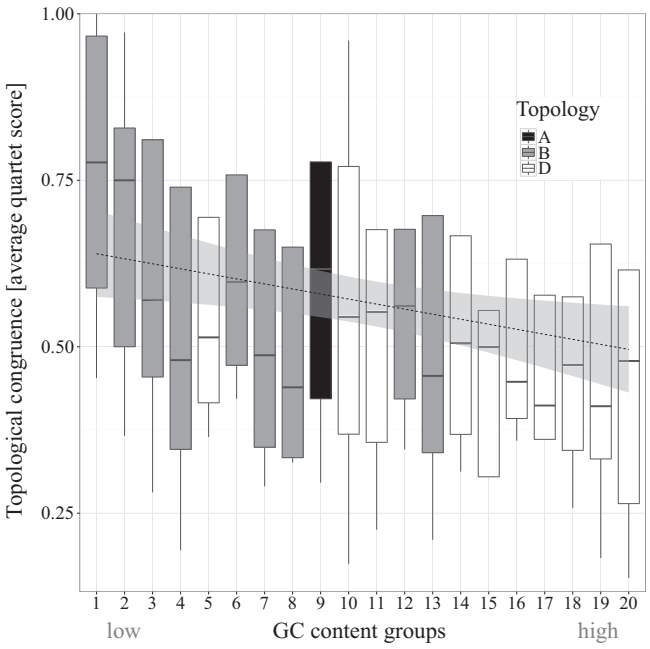

Fig. 5. General topological congruence along the trends of changing GC composition. Each content group contains 40 loci. Whiskers extend to 0.5*IQR and shaded areas indicate 95% CI.

the phylogenetic inferences. Whereas the general congruence among the gene trees decreases with increasing GC content, the summary trees of the heterogeneous loci become slightly less conflicting with higher variance of GC among taxa. In other words, increasing GC content had a negative impact on the general phylogenetic accuracy (Fig. 5). Increasing GC heterogeneity had a slightly positive but insignificant impact on phylogenetic accuracy (Supplementary data 5). The trend caused by the varying GC contents on the general congruence is stronger ($R^2 = 0.025$, $p = 0.011$) and less likely due to randomness than the grouping by heterogeneous loci ($R^2 = 0.012$, $p = 0.077$). Nonetheless, both trends are fairly weak.

Most strikingly, the GC content had strong opposing effects on the support of the three tested corbiculate topologies (Fig. 6). Loci with low GC content favor the placement of Euglossini as sister group to the remaining corbiculates ($R^2 = 0.3181$, $p = 0.009$), whereas the GC rich groups strongly support the honeybees (Apini) at this position ($R^2 = 0.4457$, $p = 0.001$). The sister-group relationship between Euglossini and Apini, which has been obtained in previous studies of corbiculate relationships based on multilocus data sets (e.g., Cardinal and Danforth, 2011; Kawakita et al., 2008; Martins et al., 2014), is overall poorly supported across a wide range of GC content with no detectable correlation between quartet support and GC content ($R^2 = 0.0261$, $p = 0.497$). Scoring the three hypotheses against the loci groups of the GC heterogeneity subset did not reveal comparable trends (Supplementary data 6). Overall, our results demonstrate that increasing GC richness has a significant negative impact on the phylogenetic accuracy obtained by using UCE data in groups such as corbiculate bees, where significant incongruence can be expected.

4. Discussion

4.1. Incongruence and GC bias as a source of error in phylogenetic analysis

Our results indicate clearly that incongruence among gene trees has played a significant role in rendering the phylogeny of corbiculate bees so challenging. Whereas the concatenation-based methods provide very strong support for our "preferred" topology (Fig. 1B), examination of the underlying gene-trees, especially along the branches involving the Apini and Euglossini, reveals significant incongruence and limited support for this topology (Fig. 2). Partitioning of loci by GC content and GC heterogeneity reveals that both GC content and (to a lesser extent) GC heterogeneity are impacting phylogenetic accuracy and incongruence among loci: GC rich loci show increasing support for the "wrong" tree topology and increased topological incongruence among loci (Figs. 5 and 6); In turn, GC heterogeneous loci provide slightly less topological incongruence but no clear trend towards recovering the "right" tree (Supplementary data 5 and 6). These results support the link between GC bias and incongruence caused by GC biased gene conversion and provide indirect evidence that gBGC is present in UCEs.

We suspect that the topological incongruence that we have documented in corbiculate bees is due largely to the unusual feature of extraordinary recombination rates in honeybees. High recombination rates are known from several eusocial insect lineages (Sirviö et al., 2011), however, the recombination rates of *A. mellifera* are significantly higher than in any other insect species for which recombination rate data are available and are considered among the highest of all higher eukaryotes (Beye et al., 2006; Wilfert et al., 2007). The variation of recombination rates along the intra-genomic landscape of the honeybee genome is further

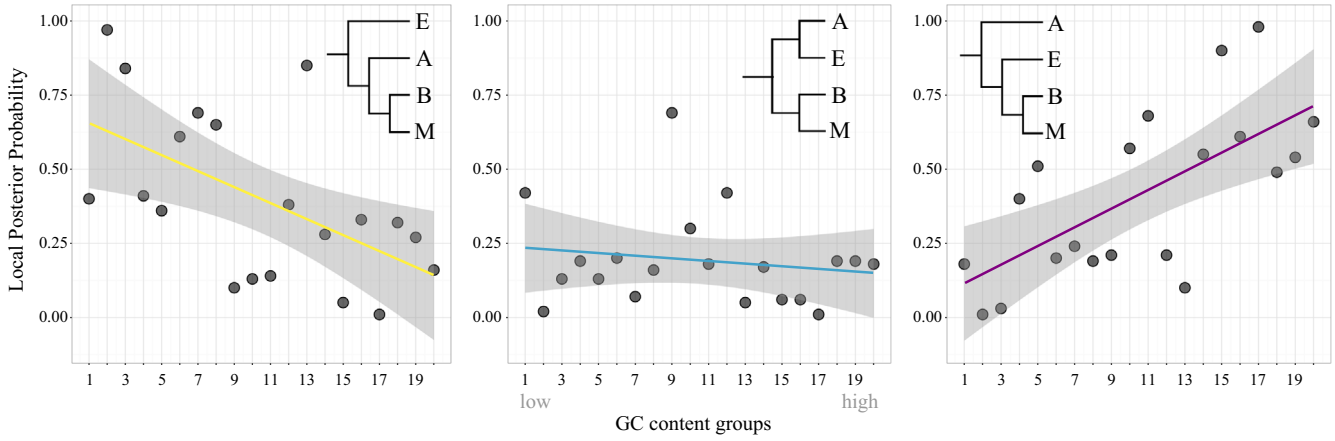

Fig. 6. Scoring the gene trees of the GC content groups against the three main hypotheses of the corbiculates relationships. Shaded areas indicate 95% CI. The regression analyses of the left and the right plot differ significantly from the null hypothesis: From left to right; (1) $R^2 = 0.318$, $p = 0.009$, (2) $R^2 = 0.026$, $p = 0.496$, (3) $R^2 = 0.445$, $p = 0.001$. Clade initials correspond to the following lineages: A – Honeybees (Apini), B – Bumblebees (Bombini), E – Orchid bees (Euglossini), and M – Stingless Bees (Meliponini).

strongly correlated to GC content (Ross et al., 2015), indicating that gBGC is an important driver of the base composition (Wallberg et al., 2015). Previous studies showed that recombination rate is positively correlated with the effective population size N_e (Hellmann et al., 2005). Together with speciation time (i.e., internode branch length), N_e in turn modulates the rate of incomplete lineage sorting, which is a major driver of gene-tree/species-tree incongruences (e.g., Degnan and Rosenberg, 2009; Maddison, 1997). This means that accumulated GC in honeybee UCEs indicates elevated recombination rates, which in turn increases the probability for ILS and subsequently higher incongruence among gene trees of the corbiculates.

This high level of incongruence presents a major challenge to determining if Euglossini or Bombini + Meliponini, is the actual sister lineage to *Apis*. Honeybees are unusual bees in other ways as well. They represent the only bee group that uses symbolic language (the dance language) to communicate the location of food resources, and they differ from other corbiculate bees in the mode of new nest founding, mating, and many other behavioral traits (Hepburn and Radloff, 2011; Michener, 1974). Studies based on a small number of genes generated with PCR and Sanger sequencing may have been particularly vulnerable to the impact of gene-gene incongruence. Even with nearly 1000 genes we find that only a small majority of genes actually supports the "correct" tree (Fig. 2). Our results provide a cautionary tale for studies of taxa in which extremely rapid diversification (short internodes) and high recombination rates can lead to substantial gene-gene incongruence. Even with massive data sets with several thousands of genes, the underlying gene tree incongruence can be obscured when only concatenation methods are used. We recommend that future studies using phylogenomic data assess incongruence on a node-by-node basis in order to identify nodes that are problematic as we have done here.

4.2. The versatility of UCEs to detect topological conflict

Because phylogenomic data sets, including those developed using targeted enrichment methods, typically include information from thousands of independent loci scattered across the genome, these methods provide an extraordinarily powerful tool for examining the impact of incongruence in phylogenetic analysis. Our case study demonstrates the importance of examining both concatenation and gene-tree methods in order to assess the level of underlying incongruence in phylogenomic data sets. Conventional concatenation based methods are more prone to yielding "incorrect" phylogenetic trees if considerable discordance is present in the examined lineages (Mirarab et al., 2014), and the respective support measures can give misleading levels of confidence (Kubatko and Degnan, 2007). Species tree methods, such as the Astral-2 algorithm, are more robust against topological conflict due to ILS than concatenation, as they account for gene tree incongruences by implementing the multi-species coalescent model (MSC; Rannala and Yang, 2003). In contrast, concatenation approaches can provide greater accuracy when data sets have low levels of ILS and horizontal gene transfer (Chou et al., 2015; Mirarab and Warnow, 2015). As the persistently inconclusive relationship of the Apini and Euglossini shows strong phylogenetic conflict, it is apparent that the coalescent-related confidence measures reflect the fragility of the inferred topology among the corbiculates more accurately than the seemingly well-supported BI and ML trees. Moreover, the exaggerated robustness of the concatenation-based confidence measures depicts their stochastic limitations, as they basically show no uncertainty despite significant underlying incongruence. This does not at all imply that they are statistically inaccurate. It does, however, reveal the vulnerability of concatenation methods to methodological shortfalls such as model misspecification when sampling genome-scale data (cf., Kumar et al., 2012). In contrast, UCEs provide superior potential to address difficult and incongruent nodes using coalescent-based gene tree methods, since the large number of individual loci provides a sufficient sample size to overcome coalescent stochasticity.

4.3. Phylogeny of the corbiculate bees – have we finally gotten this right?

Resolving the phylogeny of corbiculate bees based on molecular data has been a challenge since Sydney Cameron published the first molecular study of corbiculate relationships (Cameron, 1993; reviewed in Cardinal and Packer 2007). Previous molecular studies consistently grouped Meliponini + Bombini, but the relationships between Apini and Euglossini were less consistent with alternating topologies (cf., Cardinal and Danforth, 2011; Hedtke et al., 2013). Many studies grouped Apini + Euglossini, some with high bootstrap support and high posterior probabilities (Cardinal and Danforth, 2011, 2013; Kawakita et al., 2008). However, from todays' perspective these studies included an insufficient number of independent genes or loci to overcome the challenge of reconstructing a difficult node in the presence of significant GC bias and incongruent gene tree topologies. This topology was subsequently challenged by a 20 gene study that depicted the Euglossini on a basal branch within the corbiculates (Hedtke et al., 2013). Furthermore, the grouping of Apini + Euglossini was viewed by most morphologists as highly unlikely (Cardinal and Packer, 2007; Noll, 2002; Plant and Paulus, 2016) and also difficult to reconcile with the distribution of sociality in these four tribes. Two phylogenomic studies (Romiquier et al., 2015 and the present study) have now supported a phylogeny that places the weakly social and solitary Euglossini as sister to the eusocial tribes (Apini, Bombini, Meliponini). This basal position is also strongly supported by behavioral characters, especially in respect to the distribution of sociality (Noll, 2002), and by the phylogenetic studies based on morphology (Cardinal and Packer, 2007; Plant and Paulus, 2016). Considering the congruent evidence of morphological, behavioral, and now phylogenomic data on the placement of the Euglossini, we are confident that the remaining eusocial corbiculates form a monophyletic group. However, the sister-group relationship of Bombini and Meliponini, which is essentially inferred by every nucleotide-based phylogeny (e.g., Cameron 1993; Cardinal and Danforth, 2013; Kawakita et al., 2008; Hedtke et al., 2013) and the present study, contrasts with the topology based on morphological data (Fig. 1C).

Both phylogenomic studies detected challenges reconstructing the placement of Apini, particularly because of gene-gene incongruence. However, both studies do converge on the same topology (Fig. 2), making this our current best estimate of corbiculate relations. Nonetheless, the gene tree analyses using the multi-species coalescent model revealed two very closely competing hypothesis, which indicates the fragility of our results. Coalescent theory (see Kingman, 1982) predicts stochastically sorted genes within descendant lineages when genes fail to coalesce in ancestral populations (ILS), and hence sufficient sample sizes of independently evolving loci directly lowers the probability of false results due to randomness. Maximizing the amount of orthologous loci for gene tree based species tree estimations will therefore not solve the nature of incongruences itself but increase the confidence that the pendulum swings to the side of the true species tree.

An evident weakness of both our and Romiguier et al.'s (2015) study concerns the sparse taxon sampling, as both studies include a fairly low number of corbiculate taxa and no closely related outgroup taxon, such as *Centris* (Hedtke et al., 2013; Martins et al., 2014). Outgroup choice can impact ingroup topology (e.g., Cameron et al., 2004; Philippe et al., 2005; Ware et al., 2008) and

can even alter levels of branch support (Kirchberger et al., 2014). Strikingly, phylogenetic distance from outgroup to ingroup taxa was shown to promote incongruence among loci (Rosenfeld et al., 2012). As the accuracy of gene tree methods directly depend on the quality of the underlying input gene trees, a primary objective of future studies should be to address the optimization of the individual tree inferences and should develop more thorough taxon sampling.

Acknowledgements

We are grateful to Jason Dombroskie from the Cornell University Insect Collection (http://cuic.entomology.cornell.edu/) for the access to the high quality imaging system used to prepare Fig. 2. We further thank two anonymous reviewers for their constructive comments on the manuscript. This study was funded by a U.S. National Science Foundation grant to B.N. Danforth, S.G. Brady, J.P. Pitts, and R. Ross [DEB-1555905].

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2017.03.022.

References

Aberer, A.J., Kobert, K., Stamatakis, A., 2014. ExaBayes: massively parallel bayesian tree inference for the whole-genome era. Mol. Biol. Evol. 31, 2553–2556. http://dx.doi.org/10.1093/molbev/msu236.

Ascher, J.S., Pickering, J., 2016. Discover Life bee species guide and world checklist (Hymenoptera: Apoidea: Anthophila) <http://www.discoverlife.org/mp/20q?guide=Apoidea_species> (accessed 31.01.17).

Beye, M., Gattermeier, I., Hasselmann, M., Gempe, T., Schioett, M., Baines, J.F., Schlipalius, D., Mougel, F., Emore, C., Rueppell, O., Sirviö, A., Guzmán-Novoa, E., Hunt, G., Solignac, M., Page, R.E., 2006. Exceptionally high levels of recombination across the honey bee genome. Genome Res. 16, 1339–1344. http://dx.doi.org/10.1101/gr.5680406.

Blaimer, B.B., Brady, S.G., Schultz, T.R., Lloyd, M.W., Fisher, B.L., Ward, P.S., 2015. Phylogenomic methods outperform traditional multi-locus approaches in resolving deep evolutionary history: a case study of formicine ants. BMC Evol Biol 15, 1–14. http://dx.doi.org/10.1186/s12862-015-0552-5.

Blaimer, B.B., Lloyd, M.W., Guillory, W.X., Brady, S.G., 2016. Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. PLoS ONE 11, e0161531. http://dx.doi.org/10.1371/journal.pone.0161531.

Borowiec, M.L., 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ 4, e1660. http://dx.doi.org/10.7717/peerj.1660.

Borowiec, M.L., Lee, E.K., Chiu, J.C., Plachetzki, D.C., 2015. Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. BMC Genom. 16, 1–15. http://dx.doi.org/10.1186/s12864-015-2146-4.

Branstetter, M.G., Longino, J.T., Reyes-López, J., Schultz, T.R., Brady, S.G., 2016. Into the tropics: phylogenomics and evolutionary dynamics of a contrarian clade of ants. bioRxiv. http://dx.doi.org/10.1101/039966.

Cameron, S.A., 1993. Multiple origins of advanced eusociality in bees inferred from mitochondrial DNA sequences. PNAS 90, 8687–8691.

Cameron, S.L., Miller, K.B., D'Haese, C.A., Whiting, M.F., Barker, S.C., 2004. Mitochondrial genome data alone are not enough to unambiguously resolve the relationships of Entognatha, Insecta and Crustacea sensu lato (Arthropoda). Cladistics 20, 534–557. http://dx.doi.org/10.1111/j.1096-0031.2004.00040.x.

Cardinal, S., Danforth, B.N., 2011. The antiquity and evolutionary history of social behavior in bees. PLoS ONE 6, e21086. http://dx.doi.org/10.1371/journal.pone.0021086.

Cardinal, S., Danforth, B.N., 2013. Bees diversified in the age of eudicots. Proc. R. Soc. B 280, 1–9. http://dx.doi.org/10.1098/rspb.2012.2686.

Cardinal, S., Packer, L., 2007. Phylogenetic analysis of the corbiculate Apinae based on morphology of the sting apparatus (Hymenoptera: Apidae). Cladistics 23, 99–118. http://dx.doi.org/10.1111/j.1096-0031.2006.00137.x.

Chou, J., Gupta, A., Yaduvanshi, S., Davidson, R., Nute, M., Mirarab, S., Warnow, T., 2015. A comparative study of SVDquartets and other coalescent-based species tree estimation methods. BMC Genom. 16, 1–11. http://dx.doi.org/10.1186/1471-2164-16-s10-s2.

Danforth, B.N., Cardinal, S., Praz, C., Almeida, E.A.B., Michez, D., 2012. The impact of molecular data on our understanding of bee phylogeny and evolution. Annu. Rev. Entomol. 58, 57–78. http://dx.doi.org/10.1146/annurev-ento-120811-153633.

Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24, 332–340. http://dx.doi.org/10.1016/j.tree.2009.01.009.

Degnan, J.H., Salter, L.A., 2005. Gene tree distributions under the coalescent process. Evolution 59, 24–37. http://dx.doi.org/10.1554/04-385.

Duret, L., Galtier, N., 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genomics Hum. Genet. 10, 285–311. http://dx.doi.org/10.1146/annurev-genom-082908-150001.

Faircloth, B.C., McCormack, J., Crawford, N., Harvey, M., Brumfield, R., Glenn, T., 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61, 717–726. http://dx.doi.org/10.1093/sysbio/sys004.

Faircloth, B.C., Branstetter, M.G., White, N.D., Brady, S.G., 2015. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. Mol. Ecol. Resour. 15, 489–501. http://dx.doi.org/10.1111/1755-0998.12328.

Faircloth, B.C., 2016. PHYLUCE is a software package for the analysis of conserved genomic loci. Bioinformatics 32, 786–788. http://dx.doi.org/10.1093/bioinformatics/btv646.

Figuet, E., Ballenghien, M., Romiguier, J., Galtier, N., 2015. Biased gene conversion and GC-Content evolution in the coding sequences of reptiles and vertebrates. Genome Biol. Evol. 7, 240–250. http://dx.doi.org/10.1093/gbe/evu277.

Fontaine, M.C., Pease, J.B., Steele, A., Waterhouse, R.M., Neafsey, D.E., Sharakhov, I.V., Jiang, X., Hall, A.B., Catteruccia, F., Kakani, E., Mitchell, S.N., Wu, Y.-C., Smith, H. A., Love, R.R., Lawniczak, M.K., Slotman, M.A., Emrich, S.J., Hahn, M.W., Besansky, N.J., 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347. http://dx.doi.org/10.1126/science.1258524.

Galtier, N., Daubin, V., 2008. Dealing with incongruence in phylogenomic analyses. Proc. Roy. Soc. B: Biol. Sci. 363, 4023–4029. http://dx.doi.org/10.1098/rstb.2008.0144.

Galtier, N., Gouy, M., 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. 15, 871–879. http://dx.doi.org/10.1093/oxfordjournals.molbev.a025991.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D.B., Lander, E.S., Nusbaum, C., 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotech. 27, 182–189. http://dx.doi.org/10.1038/nbt.1523.

Goodman, M., Czelusniak, J., Moore, G.W., Romero-Herrera, A.E., Matsuda, G., 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Biol. 28, 132–163. http://dx.doi.org/10.1093/sysbio/28.2.132.

Hedtke, S., Patiny, S., Danforth, B., 2013. The bee tree of life: a supermatrix approach to apoid phylogeny and biogeography. BMC Evol. Biol. 13, 1–13. http://dx.doi.org/10.1186/1471-2148-13-138.

Hellmann, I., Prüfer, K., Ji, H., Zody, M.C., Pääbo, S., Ptak, S.E., 2005. Why do human diversity levels vary at a megabase scale? Genome Res. 15, 1222–1231. http://dx.doi.org/10.1101/gr.3461105.

Hepburn, R., Radloff, S.E. (Eds.), 2011. Honeybees of Asia. Springer, Heidelberg, Dordrecht, London, New York. http://dx.doi.org/10.1007/978-3-642-16422-4.

Hobolth, A., Dutheil, J.Y., Hawks, J., Schierup, M.H., Mailund, T., 2011. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome Res. 21, 349–356. http://dx.doi.org/10.1101/gr.114751.110.

Hodcoft, E., 2016. TreeCollapserCL 4: <http://emmahodcroft.com/TreeCollapseCL.html> (accessed 31.01.17).

Jeffroy, O., Brinkmann, H., Delsuc, F., Philippe, H., 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22, 225–231. http://dx.doi.org/10.1016/j.tig.2006.02.003.

Kapheim, K.M., Pan, H., Li, C., Salzberg, S.L., Puiu, D., Magoc, T., Robertson, H.M., Hudson, M.E., Venkat, A., Fischman, B.J., Hernandez, A., Yandell, M., Ence, D., Holt, C., Yocum, G.D., Kemp, W.P., Bosch, J., Waterhouse, R.M., Zdobnov, E.M., Stolle, E., Kraus, F.B., Helbing, S., Moritz, R.F.A., Glastad, K.M., Hunt, B.G., Goodisman, M.A.D., Hauser, F., Grimmelikhuijzen, C.J.P., Pinheiro, D.G., Nunes, F. M.F., Soares, M.P.M., Tanaka, É.D., Simões, Z.L.P., Hartfelder, K., Evans, J.D., Barribeau, S.M., Johnson, R.M., Massey, J.H., Southey, B.R., Hasselmann, M., Hamacher, D., Biewer, M., Kent, C.F., Zayed, A., Blatti, C., Sinha, S., Johnston, J.S., Hanrahan, S.J., Kocher, S.D., Wang, J., Robinson, G.E., Zhang, G., 2015. Genomic signatures of evolutionary transitions from solitary to group living. Science 348, 1139–1143. http://dx.doi.org/10.1126/science.aaa4788.

Katoh, K., Asimenos, G., Toh, H., 2009. Multiple Alignment of DNA sequences with MAFFT. In: Posada, D. (Ed.), Bioinformatics for DNA Sequence Analysis. Humana Press, Totowa, NJ, pp. 39–64.

Kawakita, A., Ascher, J.S., Sota, T., Kato, M., Roubik, D.W., 2008. Phylogenetic analysis of the corbiculate bee tribes based on 12 nuclear protein-coding genes (Hymenoptera: Apoidea: Apidae). Apidologie 39, 163–175. http://dx.doi.org/10.1051/apido:2007046.

Kent, C.F., Minaei, S., Harpur, B.A., Zayed, A., 2012. Recombination is associated with the evolution of genome structure and worker behavior in honey bees. PNAS 109, 18012–18017. http://dx.doi.org/10.1073/pnas.1208094109.

Kingman, J.F.C., 1982. The coalescent. Stoch. Proc. Appl. 13, 235–248. http://dx.doi.org/10.1016/0304-4149(82)90011-4.

Kirchberger, P.C., Sefc, K.M., Sturmbauer, C., Koblmüller, S., 2014. Outgroup effects on root position and tree topology in the AFLP phylogeny of a rapidly radiating

lineage of cichlid fish. Mol. Phylogenet. Evol. 70, 57–62. http://dx.doi.org/10.1016/j.ympev.2013.09.005.
Kjer, K.M., Simon, C., Yavorskaya, M., Beutel, R.G., 2016. Progress, pitfalls and parallel universes: a history of insect phylogenetics. J. R. Soc. Interface 13. http://dx.doi.org/10.1098/rsif.2016.0363.
Kocher, S., Li, C., Yang, W., Tan, H., Yi, S., Yang, X., Hoekstra, H., Zhang, G., Pierce, N., Yu, D., 2013. The draft genome of a socially polymorphic halictid bee, *Lasioglossum albipes*. Genome Biol. 14, R142. http://dx.doi.org/10.1186/gb-2013-14-12-r142.
Koeniger, N., Koeniger, G., Smith, D., 2011. Phylogeny of the Genus *Apis*. In: Hepburn, H.R., Radloff, S.E. (Eds.), Honeybees of Asia. Springer, Berlin, Heidelberg, pp. 23–50.
Kolaczkowski, B., Thornton, J.W., 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431, 980–984. http://dx.doi.org/10.1038/nature02917.
Kronforst, M.R., 2008. Gene flow persists millions of years after speciation in *Heliconius* butterflies. BMC Evol. Biol. 8, 1–8. http://dx.doi.org/10.1186/1471-2148-8-98.
Kubatko, L.S., Degnan, J.H., 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56, 17–24. http://dx.doi.org/10.1080/10635150601146041.
Kumar, S., Filipski, A.J., Battistuzzi, F.U., Kosakovsky Pond, S.L., Tamura, K., 2012. Statistics and truth in phylogenomics. Mol. Biol. Evol. 29, 457–472. http://dx.doi.org/10.1093/molbev/msr202.
Lartillot, N., 2013. Phylogenetic patterns of GC-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. Mol. Biol. Evol. 30, 489–502. http://dx.doi.org/10.1093/molbev/mss239.
Lemmon, A., Emme, S., Lemmon, E., 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61, 727–744. http://dx.doi.org/10.1093/sysbio/sys049.
Lemmon, E.M., Lemmon, A.R., 2013. High-throughput genomic data in systematics and phylogenetics. Annu. Rev. Ecol. Evol. Syst. 44, 99–121. http://dx.doi.org/10.1146/annurev-ecolsys-110512-135822.
Lesecque, Y., Mouchiroud, D., Duret, L., 2013. GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. Mol. Biol. Evol. http://dx.doi.org/10.1093/molbev/mst056.
Maddison, W.P., 1997. Gene trees in species trees. Syst. Biol. 46, 523–536. http://dx.doi.org/10.1093/sysbio/46.3.523.
Mancera, E., Bourgon, R., Brozzi, A., Huber, W., Steinmetz, L.M., 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature 454, 479–485. http://dx.doi.org/10.1038/nature07135.
Martins, A.C., Melo, G.A.R., Renner, S.S., 2014. The corbiculate bees arose from New World oil-collecting bees: implications for the origin of pollen baskets. Mol. Phylogenet. Evol. 80, 88–94. http://dx.doi.org/10.1016/j.ympev.2014.07.003.
McCormack, J.E., Faircloth, B.C., Crawford, N.G., Gowaty, P.A., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Res. 22, 746–754. http://dx.doi.org/10.1101/gr.125864.111.
Michener, C.D., 1974. The Social Behavior of the Bees: A Comparative Study. Harvard University Press, Cambridge, MA.
Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T., 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30, i541–i548. http://dx.doi.org/10.1093/bioinformatics/btu462.
Mirarab, S., Warnow, T., 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31, i44–i52. http://dx.doi.org/10.1093/bioinformatics/btv234.
Nater, A., Burri, R., Kawakami, T., Smeds, L., Ellegren, H., 2015. Resolving evolutionary relationships in closely related species with whole-genome sequencing data. Syst. Biol. http://dx.doi.org/10.1093/sysbio/syv045.
Noll, F.B., 2002. Behavioral phylogeny of corbiculate Apidae (Hymenoptera; Apinae), with special reference to social behavior. Cladistics 18, 137–153. http://dx.doi.org/10.1111/j.1096-0031.2002.tb00146.x.
Pamilo, P., Nei, M., 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5, 568–583. http://dx.doi.org/10.1093/oxfordjournals.molbev.a040517.
Park, D., Jung, J.W., Choi, B.-S., Jayakodi, M., Lee, J., Lim, J., Yu, Y., Choi, Y.-S., Lee, M.-L., Park, Y., Choi, I.-Y., Yang, T.-J., Edwards, O.R., Nah, G., Kwon, H.W., 2015. Uncovering the novel characteristics of Asian honey bee, *Apis cerana*, by whole genome sequencing. BMC Genom. 16, 1–16. http://dx.doi.org/10.1186/1471-2164-16-1.
Pease, J.B., Hahn, M.W., 2013. More accurate phylogenies inferred from low-recombination regions in the presence of incomplete lineage sorting. Evolution 67, 2376–2384. http://dx.doi.org/10.1111/evo.12118.
Philippe, H., Lartillot, N., Brinkmann, H., 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol. Biol. Evol. 22, 1246–1253. http://dx.doi.org/10.1093/molbev/msi111.
Plant, J.D., Paulus, H.F., 2016. Evolution and Phylogeny of Bees - a review and a cladistic analysis in light of morphological evidence (Hymenoptera, Apoidea). Zoologica 161, 1–364.
R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for statistical computing, Vienna, Austria <http://www.R-project.org/> (accessed 31.01.17).
Rannala, B., Yang, Z., 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164, 1645–1656.
Rehan, S.M., Glastad, K.M., Lawson, S.P., Hunt, B.G., 2016. The genome and methylome of a subsocial small carpenter bee, *Ceratina calcarata*. Genome Biol. Evol. 8, 1404–1410. http://dx.doi.org/10.1093/gbe/evw079.
Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804. http://dx.doi.org/10.1038/nature02053.
Romiguier, J., Cameron, S.A., Woodard, S.H., Fischman, B.J., Keller, L., Praz, C.J., 2015. Phylogenomics controlling for base compositional bias reveals a single origin of eusociality in corbiculate bees. Mol. Biol. Evol. 33, 670–678. http://dx.doi.org/10.1093/molbev/msv258.
Romiguier, J., Ranwez, V., Delsuc, F., Galtier, N., Douzery, E.J.P., 2013. Less is more in mammalian Phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. Mol. Biol. Evol. 30, 2134–2144. http://dx.doi.org/10.1093/molbev/mst116.
Rosenfeld, J.A., Payne, A., DeSalle, R., 2012. Random roots and lineage sorting. Mol. Phylogenet. Evol. 64, 12–20. http://dx.doi.org/10.1016/j.ympev.2012.02.029.
Ross, C.R., deFelice, D.S., Hunt, B.G., Ihle, K.E., Amdam, G.V., Rueppell, O., 2015. Genomic correlates of recombination rate and its variablility across eight recombination maps in the western honey bee (*Apis mellifera* L.). BMC Genom. 16, 1–11. http://dx.doi.org/10.1186/s12864-015-1281-2.
Roure, B., Philippe, H., 2011. Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. BMC Evol. Biol. 11, 1–14. http://dx.doi.org/10.1186/1471-2148-11-17.
Sadd, B.M., Barribeau, S.M., Bloch, G., de Graaf, D.C., Dearden, P., Elsik, C.G., Gadau, J., Grimmelikhuijzen, C.J., Hasselmann, M., Lozier, J.D., Robertson, H.M., Smagghe, G., Stolle, E., Van Vaerenbergh, M., Waterhouse, R.M., Bornberg-Bauer, E., Klasberg, S., Bennett, A.K., Câmara, F., Guigó, R., Hoff, K., Mariotti, M., Munoz-Torres, M., Murphy, T., Santesmasses, D., Amdam, G.V., Beckers, M., Beye, M., Biewer, M., Bitondi, M.M., Blaxter, M.L., Bourke, A.F., Brown, M.J., Buechel, S.D., Cameron, R., Cappelle, K., Carolan, J.C., Christiaens, O., Ciborowski, K.L., Clarke, D.F., Colgan, T.J., Collins, D.H., Cridge, A.G., Dalmay, T., Dreier, S., du Plessis, L., Duncan, E., Erler, S., Evans, J., Falcon, T., Flores, K., Freitas, F.C., Fuchikawa, T., Gempe, T., Hartfelder, K., Hauser, F., Helbing, S., Humann, F.C., Irvine, F., Jermiin, L.S., Johnson, C.E., Johnson, R.M., Jones, A.K., Kadowaki, T., Kidner, J.H., Koch, V., Köhler, A., Kraus, F.B., Lattorff, H.M.G., Leask, M., Lockett, G.A., Mallon, E.B., Antonio, D.S.M., Marxer, M., Meeus, I., Moritz, R.F., Nair, A., Näpflin, K., Nissen, I., Niu, J., Nunes, F.M., Oakeshott, J.G., Osborne, A., Otte, M., Pinheiro, D.G., Rossié, N., Rueppell, O., Santos, C.G., Schmid-Hempel, R., Schmitt, B.D., Schulte, C., Simões, Z.L., Soares, M.P., Swevers, L., Winnebeck, E.C., Wolschin, F., Yu, N., Zdobnov, E.M., Aqrawi, P.K., Blankenburg, K.P., Coyle, M., Francisco, L., Hernandez, A.G., Holder, M., Hudson, M.E., Jackson, L., Jayaseelan, J., Joshi, V., Kovar, C., Lee, S.L., Mata, R., Mathew, T., Newsham, I.F., Ngo, R., Okwuonu, G., Pham, C., Pu, L.-L., Saada, N., Santibanez, J., Simmons, D., Thornton, R., Venkat, A., Walden, K.K., Wu, Y.-Q., Debyser, G., Devreese, B., Asher, C., Blommaert, J., Chipman, A.D., Chittka, L., Fouks, B., Liu, J., O'Neill, M.P., Sumner, S., Puiu, D., Qu, J., Salzberg, S.L., Scherer, S.E., Muzny, D.M., Richards, S., Robinson, G.E., Gibbs, R.A., Schmid-Hempel, P., Worley, K.C., 2015. The genomes of two key bumblebee species with primitive eusocial organization. Genome Biol. 16, 1–32. http://dx.doi.org/10.1186/s13059-015-0623-3.
Sayyari, E., Mirarab, S., 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33, 1654–1668. http://dx.doi.org/10.1093/molbev/msw079.
Seo, T.-K., 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Mol. Biol. Evol. 25, 960–971. http://dx.doi.org/10.1093/molbev/msn043.
Sirviö, A., Johnston, J.S., Wenseleers, T., Pamilo, P., 2011. A high recombination rate in eusocial Hymenoptera: evidence from the common wasp *Vespula vulgaris*. BMC Genet. 12, 95. http://dx.doi.org/10.1186/1471-2156-12-95.
The Honey Bee Genome Sequencing Consortium, 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature 443, 931–949. http://dx.doi.org/10.1038/nature05260.
The Honey Bee Genome Sequencing Consortium, 2014. Finding the missing honey bee genes: lessons learned from a genome upgrade. BMC Genom. 15, 1–29. http://dx.doi.org/10.1186/1471-2164-15-86.
Wallberg, A., Glémin, S., Webster, M.T., 2015. Extreme recombination frequencies shape genome variation and evolution in the honeybee, *Apis mellifera*. PLoS Genet. 11, e1005189. http://dx.doi.org/10.1371/journal.pgen.1005189.
Wang, Y., Zhou, X., Yang, D., Rokas, A., 2015. A genome-scale investigation of incongruence in Culicidae mosquitoes. Genome Biol. Evol. http://dx.doi.org/10.1093/gbe/evv235.
Ware, J.L., Litman, J., Klass, K.-D., Spearman, L.A., 2008. Relationships among the major lineages of Dictyoptera: the effect of outgroup selection on dictyopteran tree topology. Syst. Entomol. 33, 429–450. http://dx.doi.org/10.1111/j.1365-3113.2008.00424.x.
Wheeler, W.C., Giribet, G., 2016. Molecular data in systematics: a promise fulfilled, a future beckoning. In: Williams, D., Schmitt, M., Wheeler, Q. (Eds.), The Future of Phylogenetic Systematics: The Legacy of Willi Hennig. Cambridge University Press, Cambridge, UK, pp. 329–343. http://dx.doi.org/10.1017/CBO9781316338797.015.
Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis. Springer Science +Business Media, Dordrecht, Heidelberg, London, New York. http://dx.doi.org/10.1007/978-0-387-98141-3.
Wilfert, L., Gadau, J., Schmid-Hempel, P., 2007. Variation in genomic recombination rates among animal taxa and the case of social insects. Heredity 98. http://dx.doi.org/10.1038/sj.hdy.6800950.